4-4-2019

Population Identifiability from Forensic Genetic Markers: Ancestry Variation in Latin America

Cris E. Hughes
Univeristy of Illinois Urbana Champaign, hughesc@illinois.edu

Bridget F.B. Algee-Hewitt
Stanford University, bridgeta@stanford.edu

Lyle Konigsberg
University of Illinois at Urbana-Champaign, lylek@illinois.edu

Population Identifiability from Forensic Genetic Markers: Ancestry Variation in Latin America

Cris E. Hughes,[1,3]* Bridget F.B. Algee-Hewitt,[2] and Lyle W. Konigsberg[1]

[1]Department of Anthropology, University of Illinois at Urbana–Champaign, Urbana, Illinois, USA.

[2]Center for Comparative Studies in Race and Ethnicity, Stanford University, Stanford, California, USA.

[3]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, Illinois, USA.

*Correspondence to: Cris E. Hughes, Department of Anthropology, University of Illinois at Urbana–Champaign, 607 S. Matthew Ave, Urbana, IL 61801 USA. E-mail: hughesc@illinois.edu.

Short Title: Population Identifiability from Forensic Genetic Markers

KEY WORDS: CODIS LOCI, ADMIXTURE, ANCESTRY, LATIN AMERICA, ACCURACY

Abstract

The Combined DNA Index System (CODIS) loci are a standard microsatellite marker set widely used for distinguishing among individuals in forensic DNA identity testing for medico-legal casework in the United States and in other countries. In anthropological genetic research, CODIS markers have become an important tool for uses extending beyond case investigations to quantify ancestry proportions, reveals patterns of admixture and trace population histories. These investigations are especially prevalent in studies of Latin American population structure. Nevertheless, the accuracy of the ancestry estimates computed from the CODIS loci for highly admixed Latino populations has not been formally tested. Long-standing arguments have been made that small ancestry panels, including the CODIS loci specifically, are not suitable for ancestry inference in admixed populations, due to the high heterozygosity and limited number of the loci used. Recent studies on ancestry inference using the CODIS loci suggest that these do confer more information of population-level identifiability than recognized in forensic genetic scholarship and by the medico-legal community. Here, we formally test the ability of CODIS and CODIS-Proxy (e.g. high heterozygosity and individual identifiability loci) marker panels to accurately estimate admixture proportions of individuals, including a sample of Latinos with a wide range of ancestry proportions. Using the *same* individuals in order to make direct comparisons of the outcomes, we produce ancestry estimates from 1) a small CODIS/CODIS Proxy loci panel and 2) a robust and validated microsatellite ancestry informative panel. We find evidence (e.g. $\rho = 0.80$ to 0.88) that supports the use of CODIS/CODIS-Proxy loci to capture the general ancestry estimation trends of a sample. This finding is in line with what studies using CODIS on Latin American populations have found, in that the ancestry estimations generated by CODIS present trends supported by documented population histories (e.g. colonialism and

population movements) and microevolutionary events (e.g. gene flow) in Latin America. However, the present study also highlights the limitations of CODIS for making individual-level inferences of ancestry, as the associated estimates for an acceptable level of statistical confidence (95%) are demonstrated here to be too broad to make any nuanced inferences regarding the individual's actual ancestry composition.

The Combined DNA Index System (CODIS) loci are a standard microsatellite marker set widely used for distinguishing individual identity in forensic DNA testing for medico-legal casework in the United States and in other countries. In scholarly research, CODIS loci have become an important tool beyond case investigations, particularly in research on Latin American population structure of admixture and population histories (Cerda-Flores et al., 2002; Ibarra-Rivera et al., 2008; Rubi-Castellanos et al., 2009a; Martinez-Cortes et al., 2010; Salazar-Flores et al., 2015; Rangel-Villalobos et al., 2016). Over the last decade, a plethora of published data on population variation for CODIS STRs, particularly for Mexico (Barrot et al., 2005; Sánchez et al., 2005; González-Martín et al., 2008; Rubi-Castellanos et al. 2009b; Quinto-Cortés et al., 2010; Rangel-Villalobos et al., 2013; Rangel-Villalobos et al., 2014; Martinez-Gonzalez et al., 2016) has made this research on geographic structure of genetic variations possible, with compelling inferences such as the asymmetric admixture patterns established for regions in Mexico (Rubi-Castellanos et al., 2009a). Beyond population demographic studies, the nontraditional application of CODIS loci as admixture inference-markers has also led to the recognition of ancestry-based biases in the identification process of unidentified deceased border crossers along the U.S.-Mexico border, which found that the potential of a positive identification was related to the amount of European admixture of the individual being investigated (Hughes et al., 2017).

While a steady stream of scholarly research is being produced using CODIS marker data from Latino populations, the accuracy of the CODIS panel's ancestry composition estimates for highly admixed populations has not been formally tested. Long-standing arguments have been made that CODIS loci are not suitable for ancestry inference in admixed populations, due to the high heterozygosity of these loci and limited number of loci used (Jobling and Gill, 2004; Barnholtz-Sloan et al., 2005; Silva et al., 2012). Recent findings, however, contradict these

earlier arguments (Algee-Hewitt et al. 2016), suggesting a need for closer examination of the accuracy of CODIS panels' ancestry estimations. Here, we re-examine the suitability of CODIS loci for ancestry inference, specific to the populations of Latin America, using a novel approach of identical sampling for testing the quality of forensic STR panels for ancestry inference against traditional ancestry informative panels.

Panels developed for ancestry informativeness typically include a large number of markers (hundreds to thousands), as the larger the number of markers, the greater the panel's ability to differentiate ancestral groups of interest. In contrast, ancestry informative panels with a small number of markers (< 30) have also been developed for contexts where only general continental ancestry estimates are needed, and low quantity and poor quality DNA are common, such as in forensic casework. Pardo-Seco et al. (2014) examined the stability and accuracy of small ancestry informative panels, and overwhelmingly found a positive correlation between accuracy and the number of AIMs used. While accuracy differences between thousands and hundreds of markers are trivial, the accuracy drastically differs between hundreds and only 10 markers. Pardo-Seco et al. (2014) also demonstrate that ancestry estimates for admixed individuals are more affected by smaller panels, including increased error rates associated with ancestry estimates.

The present study builds on this previous work by providing accuracy outcomes in several important ways. For example, Pardo-Seco et al.'s (2014) study focused on Asian, African and European reference samples, while the present study includes a Native American reference sample that has been demonstrated to have lower genetic diversity than other continental populations (Wang et al., 2007), and thus may impact the accuracies differently from the original study. Additionally, the present work uses forensically-relevant microsatellites, while Pardo-

Seco et al. (2014) used single nucleotide polymorphisms already targeted in ancestry informative panels. Last, Pardo-Seco and colleagues (2014) only minimally addressed the impact of admixed individuals using a simulated admixed data set all with equal contributions (33%) from each of the three reference samples. In contrast, the present study includes admixed individuals with a range of ancestry proportions.

CODIS Panel: Applications and Research. The CODIS panel was initially developed as a tool for individual identification, and has traditionally contained a suite of 13-15 "forensic" STR loci (although the U.S. standard increased to 21 loci in 2017), which were selected for characteristics that presumably enable the production of a unique genotypic profile for the individual sample, such as high observed heterozygosity (> 70%), high discriminating power (> 0.9), high level of independence or low linkage disequilibrium (LD), and ease of mixture deconvolution (Budowle et al., 1998; Butler, 2001; Hares, 2012). Since its development over 20 years ago, CODIS loci data have been incorporated into a range of applications that differ from their use for individual identification in forensic genetic profile matching. In particular, research on human genetic variation has utilized the extensive CODIS data available for analysis in public, private and federal/state databases, and also used the manufactured and validated kits for multiplex STR genotyping to amass new data. A major application of the CODIS STR variation in recent research is the estimation of ancestry proportions from forensic STRs (Pritchard et al., 2000; Wang, 2003; Alexander et al., 2009), and is well-represented in anthropological and population genetics literature for both modern and ancient populations (Barnholtz-Sloan et al., 2005; Sahoo and Kashyap 2005; Montinaro et al., 2012; Pereira et al., 2011; Phillips et al., 2011; Silva et al., 2012; Babiker et al., 2011; Mohammad et al., 2009, Callegari-Jacques et al., 2011; Rubicz et al.,

2010; Simms et al., 2010; Scliar et al., 2009; Ibarra-Rivera et al., 2008; Ricaut et at., 2005; Rubi-Castellanos et al., 2009b; Kraaijenbrink et al. 2014; Bosch et al., 2001). Additionally, more fine-grained analyses for estimating levels of admixture in individuals have also successfully been produced (Juarez-Cedillo et al., 2008; Rubi-Castellanos et al., 2009; Halder et al., 2009, Hughes et al., 2017). This breadth of work, including those studies of highly admixed individuals, implicitly speaks to the ability to use CODIS loci— selected for their ability to produce high individual identifiability—to produce information about genetic ancestry for individuals.

Algee-Hewitt et al. (2016) formally explored the relationship between individuals and population (ancestry) identifiability for the CODIS marker panel when compared to other non-CODIS marker panels (those which do not satisfy the characteristics used to select CODIS markers as described above). As they found a positive relationship between population and individual identifiability, they have provided statistical confirmation for the inherency of biogeographic ancestry information in STR loci with high individual identifiability. Using genotypes obtained for the HGDP samples, they showed specifically how markers of forensic genetic interest, notably those that make up the CODIS panel, are typically as informative as non-CODIS sets for ancestry inference. Algee-Hewitt et al.'s (2016) conclusions are remarkable in that they contradict the long-standing arguments (Jobling and Gill, 2004; Barnholtz-Sloan et al., 2005; Silva et al., 2012) that CODIS loci are not suitable for ancestry inference, due to the high heterozygosity and, so, individual identification potential of these markers. Algee-Hewitt et al., (2016) attribute these conflicting findings to the emphasis on F_{ST} as the arbitrator of ancestry information in previous studies on ancestry. They demonstrate how F_{ST} is a bounded measure that monotonically decreases as heterozygosity exceeds 0.5. They argue, accordingly, that low estimates of F_{ST} do not necessarily imply low ancestry information content for high-

heterozygosity loci – like those commonly used in forensic profile matching. This study (Algee-Hewitt et al., 2016) is an important step in understanding the utility of CODIS loci in ancestry estimations, however, because the study used HGDP samples, the results can only be directly related to individuals with minimal admixture. The present study builds on this work, by comprehensively exploring CODIS loci ancestry estimation for individuals with a range of admixture levels.

Study Objectives. Our study formally tests the ability of CODIS and CODIS-Proxy (e.g., high heterozygosity and individual identifiability loci) marker panels to accurately estimate admixture proportions of individuals. The results of our study have a direct application to those using such data to infer ancestry for both scholarly research and applied purposes. Namely, it will impact how researchers view and apply CODIS markers to study population history in Latin America. For example, to date researchers using CODIS loci report sample statistics (mean, standard deviation, etc.) for the ancestry estimates, but the actual agreement of CODIS-based ancestry estimates with validated ancestry informative panels is unknown. Statistically quantifying the relationship between CODIS and ancestry informative panels' estimates of ancestry will allow researchers using CODIS to provide reliable estimates of ancestry and error associated with their findings. The present study can capture this relationship, and the results can be integrated into future research using CODIS loci as estimators of ancestry. Finally, studying these markers' with respect to the Latino demographic is important in other fields, including forensic anthropology, which could benefit from a better understanding of alternatives uses for CODIS loci. The humanitarian crisis at the U.S.-Mexico border yields thousands of cases of deceased unidentified migrants, including workers, trafficked persons, asylum seekers and

refugees, from Latin America. Understanding genetic variation among migrant populations is essential to developing the most informed, inclusive and accurate identification protocols. CODIS STR profiles often represent the only source of genetic information available for these understudied populations in the forensic casework context, and, in addition to the skeleton, they provide another source of biological information on admixture and ancestry. Therefore, comparisons of skeletal (nonmetric and metric) and genetic (CODIS-derived) estimates of ancestry and admixture can potentially be assessed to better understand the relationship between these biological systems of data, and refine identification methods for forensic anthropologists. Furthermore, any forensic anthropology case for which CODIS STRs have been generated would be a viable study case, thus, greatly increasing the pool of available samples that can be studied for inferring the relationship between skeletal and genetic estimates of ancestry. For these reasons, we use Latin American data to evaluate the utility of forensic microsatellite markers for population inference and the estimation of admixture proportions at both continental and micro-regional scales. The dual application of CODIS markers— to generate individual identifications and admixture estimates— has important implications for anthropological genetics, population genetics, and forensic anthropology casework.

We chose to focus our analyses of admixture in Latin America for several reasons. First the majority of studies which use CODIS loci data for generating ancestry proportion estimates to consider population history are based in Latin America. Furthermore, it is important to ensure that the ability of the CODIS panel to estimate ancestry is acceptable for a range of ancestry proportions. Because ancestry estimates of admixed individuals are presumed to have more error (Pardo-Seco et al., 2014), testing CODIS in such a challenging context" (e.g., highly admixed individuals) will provide a baseline for the quality of CODIS-derived results for estimating

ancestry in the forensic setting. Latin American admixture proportions are highly heterogeneous, representing a broad range of European admixture. This heterogeneity will allow us to observe whether varying amounts of admixture impact the ability of CODIS to precisely gauge ancestry proportions.

Materials and Methods

Samples. The scope of this study requires a dataset of admixed individuals with data for genetic markers traditionally included in all CODIS loci, *as well as* validated ancestry informative markers (AIMs). Thus, the ancestry estimates generated by the CODIS panel can be compared to that of the traditional "gold standard" AIMs panel to statistically quantify their relationship and thus the accuracy of CODIS to estimate ancestry. To our knowledge, no such matched dataset exists yet. As a reasonable solution, we use the dataset described by Wang et al. (2008), which aggregates new and previously typed genotypes for 995 individuals at 678 autosomal STR markers. This dataset is adequate for the present study, in that 1) the dataset itself is comprised of admixed Latino populations from Central and South America, 2) the STR panel used was demonstrated as appropriate for accurately assessing ancestry proportions in admixed individuals with a Native American component, 3) the STR panel includes a subset of CODIS and forensically-relevant loci (reviewed below) that can represent the CODIS panel for this study. We draw from this dataset the European (n = 160) and Native American (n = 463) continental reference samples sourced from the HGDP-CEPH Human Genome Diversity Panel (Cann et al. 2002) or reported in Wang et al. (2007). We also take the full sample of admixed Latinos (n = 249 "Mestizos"), first analyzed by Wang et al. (2008). The 13 populations that

make up this sample of Latin Americans are given by geographic region and sample size in Table 1.

Genetic Marker Selection. Analyses of the of 678 markers by Wang et al. (2008), what we call the *Full Panel* in this study, revealed variation in ancestry both within and between the members of the Latino populations, even though these microsatellites were not *a priori* chosen for their informativeness of ancestry (Rosenberg 2005; Rosenberg et al. 2003). Given this prior work, we let the individual proportions of ancestry inferred for the Latino sample using the *Full Panel* of loci serve as a gold standard against which ancestry proportions estimated from smaller CODIS STR panel can be compared.

The *Full Panel* dataset of 678 STRs from Wang et al. (2008) contains five autosomal STR markers prominent in forensic analyses (D13S17, D16S539, D19S433, D7S820, D6S1043) evidenced by their inclusion in the core, now expanded, CODIS loci panel, and/or included in multiplex systems traditionally used in forensic human identification applications (e.g., PowerPlex® 21 System) (Budowle et al. 1998; Budowle et al. 2001; Butler 2005; Butler 2006; Butler 2010; Butler and Hill 2012; Butler et al. 2003; Gill 2002; Hares 2012a; Hares 2012b; Hares 2015; Schneider 2009). This subset of forensic identification loci is the largest available in any of the suitably sized , multi-populational and public-access datasets of random markers (Pemberton et al. 2009). Furthermore, the high coverage of this Wang et al. (2008) dataset also includes additional loci with forensic genetic marker properties, which, when added to the preexisting five forensic loci, allow for the creation of 15-STR *CODIS-Proxy Panels* composed of CODIS and CODIS-like markers. Since a complete CODIS-specific panel is unavailable, these 15-STR *CODIS-Proxy Panels* are used to approximate CODIS's performance in the

present study. The non-CODIS microsatellites that make up our pool of potential *CODIS-Proxy* loci are widely separated, highly polymorphic and typically not associated with a known coding gene. Therefore, we expect these, like even more closely spaced markers, to display pairwise-independence, with insignificant linkage disequilibrium (Rosenberg and Calabrese, 2004; Ghebranious et al., 2003)

Selecting STRs for the CODIS-Proxy Panels. To create the 15-STR *CODIS-Proxy Panels* required for ancestry estimation, we identified candidate STR loci within the *Full Panel* dataset using criteria that are known to define the CODIS loci and are said to characterize markers suitable for individual identification in the forensic context (Algee-Hewitt et al. 2016; Budlowe et al, 1998; Butler et al 2001; Hares 2012; Butler 2006). Given the almost exclusive use of tetranucleotide STR markers in human identification practice (Butler, 2006; Hares, 2012a; Hares, 2012b; Phillips, 2013; Hares, 2015), all of the microsatellites classified as penta-, tri- and dinucleotides were removed from consideration to be included in the *CODIS-Proxy Panels* (Pemberton et al, 2009, 2013). Next we considered the heterozygosity of the remaining loci, and our approach draws specifically on the recent work by Algee-Hewitt et al. (2016). These authors demonstrated that the CODIS loci are especially good for individual identification because they have greater heterozygosity (mean H = 0.796) and lower match probability (mean $M = 0.074$) than randomly selected sets of non-CODIS tetranucleotides. They also reported that these two criteria, H and M, are strongly inversely correlated ($r = -0.97$). These findings suggest that our *CODIS-Proxy* markers can be selected on the size of their estimated values of heterozygosity, H. We set our threshold for H to > 0.7, which, while less than the mean value reported in Algee-

Hewitt et al. (2016), is in agreement with expectations for the original set of CODIS loci in the forensic genetic literature (Butler 2001).

We, thus, calculated the value of *H* for all remaining STRs, and based this calculation on the three continental reference samples (Native American, African, and European) available in Wang et al. (2008), reasoning that analyses of heterezyogosity for STRs being considered for forensic purposes traditionally include a multicontinental sample. Any STRs with $H < 0.7$ were removed from consideration to be included in the *CODIS-Proxy Panels*. We used the remaining 199 STR loci as the pool of "forensically relevant" markers, termed our *CODIS-Proxy* STRs. We then created panels of 15 STRs which we call the *CODIS-Proxy Panels.* Each *CODIS-Proxy Panels* includes the same five forensically significant STRs (D13S17, D16S539, D19S433, D7S820, D6S1043), and the remaining 10 STRs were randomly drawn from the 199 "forensically relevant" STR loci. None of the randomly drawn STRs were duplicated in any of the 10 panels. The STRs comprising each of these *CODIS-Proxy Panels* are given in Table 2.

Estimating Ancestry Proportions. We performed supervised model-based clustering on all 11 datasets (the *Full Panel* and the 10 *CODIS-Proxy Panels*) with the program STRUCTURE 2.3.4 In order to allow for maximum comparability between results, we opted to adhere closely to the approach delineated in Wang et al. (2008). We employed, therefore, an admixture model with correlated allele frequencies, specifying identical parameters for each implementation. A supervised approach to the analysis was performed, such that individuals from reference population samples were assigned to *K* predetermined clusters. Because Wang et al. (2008) demonstrated that the African contribution to the Latino samples was consistently low, with ancestry estimates <10% across the 13 subsamples assayed, our STRUCTURE analyses included

only two reference samples (Native American and European), and we pre-specified the number of clusters, so K = 2. By imposing this two-cluster model, we assumed that this solution would produce components that align with the Native American and European samples, which represent the continental ancestries most relevant to our analyses of the Latino populations available in Wang et al. (2008). We also held the individuals constant across all STRUCTURE runs, regardless of changes to the composition of the Panel, i.e., the number or choice of markers. Thus, the ancestry estimates produced for the Full Panel and CODIS-Proxy Panel are directly comparable for assessing their correspondence.

To produce a single set of ancestry estimates for the *Full Panel* and each of the *CODIS-Proxy Panels*, we used CLUMPP 1.1.2. to compile the multiple STRUCTURE output files resulting from 10 replicate STRUCTURE 2.3.4 runs. These consensus Panels are used in an all subsequent analyses. For this *CODIS-Proxy Panels,* we visualized the patterns of ancestry by plotting for each sampled individual their fraction of membership across the two inferred clusters of European and Native American ancestry components (Rosenberg 2004). Owing to the two-cluster model, the European and Native American coefficients sum to 1.0 for both the *Full* and *CODIS-Proxy Panel* datasets such that Native American ancestry estimates increase just as European ancestry estimates decrease. Thus, when reporting analyses on these estimates, only a single vector of posterior probabilities is discussed.

Tests of Differences. To evaluate how ancestry proportions differ by the choice of markers, we calculated, for each individual, the differences between European ancestry as estimated by the *Full Panel* and the *CODIS-Proxy Panels*. The differences were plotted to reveal patterns in the differences across population and panel.

Test of Linear Relationship and Individual Predictions of Ancestry for Unknown Cases.

Spearman rank correlation analysis was used to evaluate the magnitude and direction of the association between the membership coefficients for the European ancestry component produced from each of the *Full* and *CODIS-Proxy Panels* (Chen and Popovich 2002). Because the data (ancestry proportions) of interest is probability data and thus constrained between 0 and 1, we converted all estimates of ancestry proportions to a standard normal (probit) scale. Spearman rank correlation coefficients, ρ, were calculated using the scaled European cluster membership obtained with the *Full* and *CODIS-Proxy Panels* for the pooled sample of Latino, European and Native American individuals. The statistical significance of all correlations was determined by testing if $\rho = 0$ at $\alpha = 0.05$.

While the majority of studies using CODIS markers are making inferences of individual ancestry estimates at a population level, it is pertinent to understand the suitability of CODIS markers to predict the "gold standard" ancestry for an unknown individual (e.g., a new observation). Because the analysis using Wang et al.'s approach is based on a large number of STRs, the estimates of ancestry could be considered as a "gold standard." In contrast, ancestry estimates based on the *CODIS-Proxy Panels* with a smaller set of 15 markers are easier to obtain but are generally less accurate. This consequently places our analysis within a calibration setting (Brown 1993). In this setting one can use "inverse calibration" where the Wang et al. estimates are regressed onto estimates from 15 STRs, or one can use "classical calibration" where the 15 STR estimates are regressed onto the Wang et al. estimates. In the latter case, the regression is "inverted" by solving the regression for Wang et al.'s estimates. Classical calibration is generally preferred (Chow and Shao 1990; Krutchkoff 1967; Krutchkoff 1969) because it avoids

the problem of overestimating Native American ancestry for those below the mean Native American ancestry and underestimating Native American ancestry for those above the mean Native American ancestry. Letting x represent the estimates from Wang et al.'s markers and y represent the estimates from a 15 STR panel, the initial regression is:

$$y = \alpha + \beta x. \quad (1)$$

Solving for x gives:

$$x = -\frac{\alpha}{\beta} + \frac{1}{\beta} y. \quad (2)$$

Note that equation (2) is written in the same form as a usual linear regression, so that the first term is an intercept and the second term (the multiplier for y) is a slope. Equation (2) is consequently very easy to apply.

In addition to wanting a point estimate of ancestry, we also want individual estimates of the prediction interval or credible interval for ancestry. This is a more complicated problem that has been dealt with by a number of authors (Freund and Wilson 1998:65-67; Montgomery and Peck 1982:400-405; Montgomery et al. 2006:488-489; Neter et al. 1985:172-174; Neter et al. 1990:173-176; Seber and Lee 2003:146; Snedecor and Cochran 1989:170-172; Sprent 1969:97-99; Zar 1984:276-278). Their method is to construct prediction intervals for the regression of y on x, and then solve for the values of x on the prediction intervals that coincide with the observed value for y. If the sample size is small then the prediction intervals obtained this way are generally asymmetric around the estimate. But if the sample size is large, which it is in this case, the credible intervals for x are symmetric. Further, it is easy to calculate an asymptotic posterior variance for the estimate. Hunter and Lamboy (1981) give an approximation to the posterior variance on their page 326 which is:

$$\frac{(N+1)s^2_{y|x}}{N\beta^2_{y|x}}, \tag{3}$$

where $s^2_{y|x} = \left[\text{var}(y) - \beta_{y|x} \times \text{cov}(x,y)\right] \times (N-1)/(N-2)$ and $\beta_{y|x} = \text{cov}(x,y)/\text{var}(x)$. This is still intended as an approximation for a small sample. But as N increases equation (3) approaches:

$$\text{var}(x)\left(r^{-2} - 1\right), \tag{4}$$

which was given in Table 3 of Konigsberg et al. (1998).

To show that equation (4) gives results very close to the more complicated method of inverting prediction intervals, Figure 3 compares the two methods for the *Panel* 7. This panel has the most missing data, and consequently gives the smallest sample size (N=731) of any of the four panels. A problem with the analysis as presented is that admixture estimates are not constrained to be between 0 and 1. This can be addressed by working in a standard normal (probit) scale and then converting back to the original admixture scale. The use of a probit scale does complicate the analysis in that the distribution of admixture estimates on the original scale are no longer normally distributed. As a consequence, the distributions need to be integrated and divided by the integral in order to find the highest posterior densities.

Cross-Classification & Matching Accuracy. While the previous analyses focus on how the proportions of ancestry vary with the properties of the two Panels, these kinds of tests do not tell us about the consequences that these differences in membership components have on inferring the major ancestral contributor. For example, while there may be a difference of 15% in the estimation of European ancestry between the two Panels, does this make a difference in the

hard cluster assignments (e.g. major ancestry contribution) of the individual to a particular cluster? Therefore, we investigate if the observed differences in the posterior probabilities of component membership between the *Full* and *CODIS-Proxy Panels* are of sufficient magnitude to effect differences in the hard cluster assignments that are subsequently produced from these data. We selected to run this analysis on only two of the 10 *CODIS-Proxy Panels,* selecting those Panels which had the strongest (*CODIS-Proxy Panel* 7) and weakest (*CODIS-Proxy Panel* 8) correlations with the *Full Panel.* Hard clustering was performed by assigning each individual into one of the two inferred components, corresponding to either European (k1) or Indigenous (k2) ancestry, based on the highest posterior probability of k-cluster membership. To evaluate the relationship between these hard-cluster assignments obtained with the *Proxy versus Full Panels*, cross-classification was performed (Kohavi and Provost 1998), taking the cluster assignments inferred by the *Full Panel* as the gold standard for such estimation and, so, the true memberships for the purpose of evaluating rates of classification accuracy. From the cross-classification results, we computed the match error statistic (%), defined simply as the frequency with which individuals classified by the *Full Panel* dataset as either European or Indigenous were *not* similarly classified as European or Indigenous by the hard-cluster assignments derived from the *CODIS-Proxy Panel*. Chi-Square (X^2) test was used, when appropriate, to identify a statistically significant relationship between the two sets of hard-cluster assignments at $\alpha = 0.05$.

Results

Generation of Ancestry Estimates. Under the preferred model of $K = 2$, STRUCTURE runs for both the *Full* and *CODIS-Proxy Panels* produced supervised clusters that, as expected,

corresponded with European and Native American population affinity, respectively. Individual ancestry proportions for this $K = 2$ model are displayed in Figure 1 for the both the continental reference samples and the Latino population. STRUCTURE produces information on the percent of missing loci per individual for each Panel, and was used to exclude individuals with excessively missing data (here we define that as 10% missing) that may bias analyses. For each Panel, any individual missing more than 10% of the markers included in that panel were removed from the samples for the following analyses, and modified sample sizes are included when pertinent.

When observing the individual posterior probabilities of cluster membership produced by both the *Full* and *CODIS-Proxy Panels,* we see that hard cluster classifications (defined by a posterior probability > 0.50) for 97-100% of the European sample allocate to the same cluster. Additionally, the *Full and CODIS-Proxy Panels* produced hard cluster classifications for 100% and 86-94%, respectively, of the Native American sample to a single cluster. These clustering trends allow us to assume that the posterior probabilities associated to the two clusters can be inferred as an indigenous (e.g. Native American) and non-indigenous (e.g. European or the admixture cluster), although a small component of the Native American cluster likely includes non-European admixture associated with African variation (Wang et al., 2008). Thus, the matrix of individual posterior probabilities of membership in the two inferred clusters are interpreted here as estimates of European and Native American ancestry. As expected, the individuals comprising the Native American reference sample on average exhibit minimal admixture ($\mu_{Full\ Panel} = 0.10$), and the Latino sample on average exhibits a larger amount of admixture ($\mu_{Full\ Panel} = 0.56$) as compared to the Native American reference sample. These trends are consistent across

both the *Full and CODIS-Proxy Panels*, suggesting their general agreement in ancestry estimations.

Tests of Differences. Figure 2 presents the box plots of the individual differences in percent ancestry estimates for the *Full Panel* with each of the 10 *CODIS-Proxy Panels*, with positive values indicating the *CODIS-Proxy Panel* underestimates European ancestry for a given individual as compared with the Full Panel, while negative values indicate the *CODIS-Proxy Panel* over-estimates European ancestry. General trends in Figure 2 indicate that both European and Native American ancestry are being underestimated in Europeans and Native Americans, respectively.

Test of Linear Relationship and Individual Predictions of Ancestry for Unknown Cases.

Table 3 gives the number of cases for each panel, the correlation between Wang et al.'s estimates and the panel estimates on the probit scale, and the intercept, slope, and posterior standard deviation all on the probit scale. Significant positive correlations were found between ancestry component estimates of the *Full* and *CODIS-Proxy Panels* for the Latino sample, as well as the total pooled sample (European and Native American reference samples and the Latino sample), and are presented. While the correlations are robust across the panels (0.81-0.88), the predictive relationships between the *Full* and *CODIS-Proxy Panels* highlight the error associated with this relationship. Recall that only Panel 7 was used to produce the regression. Figure 4a provides the widths of the 95% confidence interval for predicting the "gold standard" Native American ancestry response given the estimate of Native American ancestry from *CODIS-Proxy Panel* 7. In addition, Figure 4b provides the widths of the 95% highest posterior

density (HPD) for an unknown individual's "gold standard" Native American ancestry, given the estimate of Native American ancestry from *CODIS-Proxy Panel* 7. Both the confidence interval and the HPD widths are quite large, given that the full range for an ancestry estimate is 0-1.00. For example, Figure 5 shows the 95% HPD for a *CODIS-Proxy Panel* estimate of 50% Native American admixture. The estimated "gold standard" value is 51.84%, close to the 50% value from the *CODIS-Proxy Panel*, but the range for the 95% HPD, which accounts for error, is quite substantial, running from 11.54% to 90.16%. For a *CODIS-Proxy Panel* estimate of 10% Native American Admixture, the estimated Wang et al. value is 4.68% and again the 95% highest posterior density is quite large (from 0.00% to 51.74%). Finally, with a *CODIS-Proxy Panels* estimate of 90%, the estimated Wang et al. value is 96.15% with a 95% highest posterior density from 50.47% to 100.00%.

Cross-Classification and Matching Accuracy. Percent match errors, or rates of disagreement, were calculated to test if the fluctuations in ancestry proportions estimated by the *Full* and *CODIS-Proxy Panels* produce changes to the hard-cluster allocations for the individual. Table 4 gives the percent match errors for the European and Native American reference samples, and the Latino sample. The results of Chi-Square testing for the cross-classifications were significant at $\alpha = 0.05$ for the Latino sample for *CODIS-Proxy Panel* 7 (R^2= 0.15, df=1, X^2= 43.06, $Prob > X^2$= <0.0001) and CODIS-Proxy Panel 8 (R^2= 0.27, df=1, X^2= 42.53, $Prob > X^2$= <0.0001). The Native American and European reference samples were excluded from Chi-Square analysis as all of the hard-clustering classifications produced 100% assignment of these individuals to their respective cluster.

Discussion

There is clear evidence for a relationship between the ancestry estimations produced by the *CODIS-Proxy Panels* and the *Full Panel*. This is manifested in the general agreement between the STRUCTURE plots (Figure 1) and the statistically significant positive correlations (Table 3) between all 10 *CODIS-Proxy Panels* and the *Full Panel*. However, the strength of this relationship does depend on the population in question, as evidenced by the analysis of differences (Figure 2). The box plots exhibit a trend, where the differences for the reference samples are closest to zero, and trend slightly positive for the European reference sample (differences range from 0.03 to 0.083 across *Panels* 1-10), and slightly negative for the Native American reference sample (mean differences range from -0.15 to -0.05 across *Panels* 1-10). From the perspective of expectations for variability in cluster-derived values, any difference in cluster membership smaller than 10% can potentially be due to variance instead of actual differences (Phillips 2015). Therefore, the differences in admixture estimates produced by the *Full* and *CODIS-Proxy Panels* for the two reference panels are generally not notable. We see more substantial negative values for the differences for the Latino sample, with mean Panel differences ranging from -0.25 to -0.16, although on Panel's (Panel 7) mean difference was considerably less, at -0.05. This increase in differences for the Latino sample suggests that for the *CODIS-Proxy Panel* ancestry estimates are less accurate for admixed individuals than individuals with minimal admixture.

These trends towards underestimating the primary ancestry for minimally admixed individuals (as seen with our reference samples) can be interpreted as products of the panels themselves, and their ability to capture variation between clusters (Pardo-Seco et al., 2014). For example, the *Full Panel* is comprised of a large number of ancestry-informative loci, and thus

will produce K=2 clusters (European and Native American clusters) in STRUCTURE that are less overlapping due to their increased ability to capture the variation between the clusters. In contrast, the *CODIS-Proxy Panels*, comprised of only 15 loci will not capture as much of the two reference groups' variation and thus produce less distinct, more overlapping clusters. This in effect, will render posterior probabilities that are more evenly distributed between the two clusters for the STRUCTURE analyses based on the *CODIS-Proxy Panels*. This will produce the patterns observed in Figure 2, where individuals expected to have large posterior probabilities associated with a single cluster (e.g. the European and Native American reference samples), will consistently share a greater component of that posterior probability with the second available cluster. Thus, we see that Native American individuals tend to have their Native American ancestry underestimated (and their European ancestry overestimated), while European individuals tend to have their European ancestry underestimated (and their Native American ancestry overestimated).

Beyond the mean differences present in Figure 2, the range of differences is also noteworthy, as it indicates that the accuracy of the *CODIS-Proxy Panels* can greatly vary. If these panels were better estimators of ancestry, we would expect to see the spread of the differences to be much smaller. The deviations of the *CODIS-Proxy Panels* for admixed individuals are comparable to the small AIMs panels tested by others (e.g., Pardo-Seco et al., 2014). The *CODIS-Proxy Panels* appear to outperform the tested 10 AIM panel (Lao et al., 2006), are on par with the test 23 AIM panel (Corach et al. 2010), but fail to reach the smaller error rates associated with the remaining tested panels. Even when the comparisons of ancestry between the *Full* and *CODIS-Proxy Panels* are distilled down to hard cluster assignments, there are still significant deviations at the individual levels, as evidenced by the matched pairs results

for the admixed individuals (Table 4). In the Latino sample, approximately 58% of the sample (for *CODIS-Proxy Panel* 7) was assigned to the incorrect cluster, while match error rates are much lower for the two reference samples (1.34 - 4.21%). This extreme difference in match error rates is presumably a result of the Latino sample encompassing admixed individuals, whose ancestry proportions are closer to the cluster assignment threshold of 0.50, and thus more likely to produce a match error when comparing the two panels.

Finally, we produced the linear regression and associated 95% confidence and HPD intervals suggest that individual predictions of "gold standard" ancestry from *CODIS-Proxy Panels*. The ranges of HPD, regardless of the ancestry proportions of a given individual, are so wide as to render them useless in both the forensic context for individual ancestry predictions. Because most researchers are making population-level, not individual-level inferences, we provide the 95% confidence intervals for the mean response associated with the model. These results are useful references for researchers reporting and analyzing CODIS-based estimates of ancestry.

While the present study uses proxy panels to capture the expected trends of the actual CODIS panel, there is no reason to expect that the actual CODIS would outperform the present proxy panels. If anything, because the STRs used for the present study were developed to estimate ancestry on admixed populations with Native American components, one could argue that the proxy panels here are potentially better estimators of ancestry for Latino populations than the actual CODIS panel. Based on the results of this study, we find evidence (e.g. $\rho = 0.80$ to 0.88) that supports the use of CODIS to capture the general ancestry estimation trends of a sample. This finding is in line with what studies using CODIS on Latin American populations have found, in that the ancestry estimations generated by CODIS present trends supported by

documented population history trends (e.g., colonialism and population movements) and microevolutionary events (e.g., gene flow) in Latin America (Cerda-Flores et al., 2002; Ibarra-Rivera et al., 2008; Rubi-Castellanos et al., 2009a; Martinez-Cortes et al., 2010; Salazar-Flores et al., 2015; Rangel-Villalobos et al., 2016; Hughes et al., 2016). However, the present study also highlights the limitations of CODIS for making individual-level inferences of ancestry, as the associated estimates for an acceptable level of statistical confidence (95%) are demonstrated here to be too broad to make any nuanced inferences regarding the individual's actual ancestry composition.

Received 31 October 2018; accepted for publication 26 February 2019.

Literature Cited

Alexander, D. H., J. Novembre, K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19:1,655–1,664.

Algee-Hewitt, B. F. B., M. D. Edge, J. Kim et al. 2016. Individual identifiability predicts population identifiability in forensic microsatellite markers. *Curr. Biol.* 26:935–942.

Atkinson, A. C. 1985. *Plots, Transformations, and Regression.* New York: Oxford University Press.

Babiker, H. M., C. M. Schlebusch, H. Y. Hassan et al. 2011. Genetic variation and population structure of Sudanese populations as indicated by 15 Identifiler sequence-tagged repeat (STR) loci. *Investig. Genet.* 2:12.

Barnholtz-Sloan, J. S., C. L. Pfaff, R. Chakraborty et al. 2005. Informativeness of the CODIS STR loci for admixture analysis**.** *J. Forensic Sci.* 50:1,322–1,326.

Barrot, C., C. Sánchez, M. Ortega et al. 2005. Characterisation of three Amerindian populations from Hidalgo State (Mexico) by 15 STR-PCR polymorphisms. *Int. J. Legal Med.* 119:111–115.

Bosch, E., J. Clarimón, A. Pérez-Lezaun et al. 2001. STR data for 21 loci in northwest Africa. *Forensic Sci. Int.* 16:41–51.

Brown, P. J. 1993. *Measurement, Regression, and Calibration.* New York: Oxford University Press.

Budowle, B., T. R. Moretti, S. J. Niezgoda et al. 1998. CODIS and PCR-based short tandem repeat loci: Law enforcement tools. *Second European Symposium on Human Identification,* Madison: Promega Corporation, 73–88.

Budowle, B., B. Shea, S. Niezgoda et al. 2001. CODIS STR loci data from 41 sample populations. *J. Forensic Sci.* 46:453–489.

Butler, J. M. 2001. *Forensic DNA Typing.* London: Academic Press.

Butler, J. M., R. Schoske, P. M. Vallone et al. 2003. Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic Populations. *J. Forensic Sci.* 48:908–911.

Butler, J. M. 2005. *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers.* New York: Elsevier.

Butler, J. M. 2006. Genetics and genomics of core short tandem repeat loci used in human identity testing. *J. Forensic Sci.* 5:253–265.

Butler, J. M. 2010. *Fundamentals of Forensic DNA Typing.* Boston, MA: Academic Press.

Butler, J. M., and C. R. Hill. 2012. Biology and genetics of new autosomal STR loci useful for forensic DNA analysis. *Forensic Sci. Rev.* 24:15–26.

Callegari-Jacques, S. M., E. M. Tarazona-Santos, R. H. Gilman et al. 2011. Autosome STRs in native South America—testing models of association with geography and language. *Am. J. Phys. Anthropol.* 145:371–381.

Cann, H. M., C. de Toma, L. Cazes et al. 2002. A human genome diversity cell line panel. *Science* 296:261–262.

Cerda-Flores, R., B. Budowle, L. Jin et al. 2002. Maximum likelihood estimates of admixture in northeastern Mexico using 13 short tandem repeat loci. *Am. J. Hum. Biol.* 14:429–439.

Chen, P. Y., and P. M. Popovich. 2002. *Correlation: Parametric and Nonparametric Measures.* Thousand Oaks, CA: SAGE Publications.

Chow, S.-C., and J. Shao. 1990. On the difference between the classical and inverse methods of calibration. *J. R. Stat. Soc. Ser. C Appl. Stat.* 39:219–228.

Corach, D., O. Lao, and C. Bobillo. 2010. Inferring continental ancestry of Argentineans from autosomal, Y-chromosomal and mitochondrial DNA. *Ann. Hum. Genet.* 74: 65–76.

Freund, R. J., and W. J. Wilson. 1998. *Regression Analysis: Statistical Modeling of a Response Variable.* San Diego, CA: Academic Press.

Ghebranious, N., D. Vaske, A. Yu et al. 2003. STRP screening sets for the human genome at 5 cM density. *BMC Genomics* 4:6.

Gill, P. 2002. Role of short tandem repeat DNA in forensic casework in the UK--past, present, and future perspectives. *Biotechniques* 32:366–385.

González-Martín, A., A. Gorostiza, H. Rangel-Villalobos et al. 2008. Analyzing the genetic structure of the Tepehua in relation to other neighbouring Mesoamerican populations. A study based on allele frequencies of STR markers. *Am. J. Hum. Biol.* 20:605–613.

Halder, I., B. Z. Yang, H. R. Kranzler et al. 2009. Measurement of admixture proportions and description of admixture structure in different US populations. *Hum. Mutat.* 30:1,299–1,309.

Hares, D. R. 2012a. Addendum to expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genetics* 6:e135.

Hares, D. R. 2012b. Expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genetics* 6:e52–e54.

Hares, D. R. 2015. Selection and implementation of expanded CODIS core loci in the United States. *Forensic Sci. Int. Genetics* 17:33–34.

Hughes, C. E., B. F. B. Algee-Hewitt, R. Reineke et al. 2017. Temporal patterns of Mexican migrant genetic ancestry: Implications for identification. *Am. Anthropol.* 119:193–208.

Hunter, W. G., and W. F. Lamboy. 1981. A Bayesian analysis of the linear calibration problem. *Technometrics* 23:323–350.

Ibarra-Rivera, L., S. Mirabal, M. M. Regueiro et al. 2008. Delineating genetic relationships among the Maya. *Am. J. Phys. Anthropol.* 135:329–347.

Jakobsson, M., and N. A. Rosenberg. 2007. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* 23:1,801–1,806.

Jobling, M. A., and P. Gill. 2004. Encoded evidence: DNA in forensic analysis**.** *Nat. Rev. Genet.* 5:739–751.

Juárez-Cedillo, T., J. Zuñiga, V. Acuña-Alonzo et al. 2008. Genetic admixture and diversity estimations in the Mexican Mestizo population from Mexico City using 15 STR polymorphic markers. *Forensic Sci. Int. Genetics* 2*:*e37–e39.

Kohavi, R., and F. Provost. 1998. Guest editors' introduction: On applied research in machine learning. *Mach. Learn.* 30:127–132.

Konigsberg, L. W., S. M. Hens, L. M. Jantz. 1998. Stature estimation and calibration: Bayesian and maximum likelihood perspectives in physical anthropology. *Yearb. Phys. Anthropol.* 41:65–92.

Kopelman, N. M., J. Mayzel, M. Jakobsson et al. 2015. Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* 15:1,179–1,191.

Kraaijenbrink, T., K. J. van der Gaag, S. B. Zuniga et al. 2014. A linguistically informed autosomal STR survey of human populations residing in the greater Himalayan region. *PloS One* 9:e91534.

Krutchkoff, R. G. 1967. Classical and inverse regression methods of calibration. *Technometrics* 9:425–439.

Krutchkoff, R. G. 1969. Classical and inverse regression methods of calibration in extrapolation. *Technometrics* 11:605–608.

Lao, O., K. van Duijn, P. Kersbergen et al. 2006. Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am. J. Hum. Genet.* 78:680–690.

Martínez-Cortés, G., I. Nuño-Arana, I., R. Rubi-Castellanos et al. 2010. Origin and genetic differentiation of three Native Mexican groups (Purépechas, Triquis and Mayas): Contribution of CODIS-STRs to the history of human populations of Mesoamerica. *Ann. Hum. Biol.* 37:801–819.

Martinez-Gonzalez, L. J., M. J. Alvarez-Cubero, M. Saiz et al. 2016. Characterisation of genetic structure of the Mayan population in Guatemala by autosomal STR analysis. *Ann. Hum. Biol.* 43:457–468.

Mohammad, T., Y. Xue, M. Evison et al. 2009. Genetic structure of nomadic Bedouin from Kuwait. *Heredity (Edinb)* 103:425.

Montgomery, D. C., and E. A. Peck. 1982. *Introduction to Linear Regression Analysis.* New York: Wiley.

Montgomery, D. C., E. A. Peck, and G. G. Vining. 2006. *Introduction to Linear Regression Analysis.* Hoboken, NJ: Wiley-Interscience.

Montinaro, F., I. Boschi, F. Trombetta et al. 2012. Using forensic microsatellites to decipher the genetic structure of linguistic and geographic isolates: A survey in the eastern Italian Alps. *Forensic Sci. Int. Genetics* 6:827–833.

Neter, J., W. Wasserman, and M. Kutner. 1985. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs.* Homewood, IL: Irwin.

Neter, J., W. Wasserman, and M. Kutner. 1990. *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs.* Homewood, IL: Irwin.

Pardo-Seco, J., F. Martinón-Torres, and A. Salas. 2014. Evaluating the accuracy of AIM panels at quantifying genome ancestry. *BMC Genomics* 15:543.

Pemberton, T. J., C. I. Sandefur, M. Jakobsson et al. 2009. Sequence determinants of human microsatellite variability. *BMC Genomics* 10:1–19.

Pemberton, T. J., M. DeGiorgio, and N. A. Rosenberg. 2013. Population structure in a comprehensive genomic data set on human microsatellite variation. *G3 (Bethesda)* 8:3–113.

Pereira, L., F. Alshamali, R. Andreassen et al. 2011. PopAffiliator: Online calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile. *Int. J. Legal Med.* 125:629–636.

Phillips, C., L. Fernandez-Formoso, and M. Garcia-Magarinos. 2011. Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. *Forensic Sci. Int. Genetics* 5:155–169.

Phillips, C., L. Fernandez-Formoso, M. Gelabert-Besada et al. 2013. Development of a novel forensic STR multiplex for ancestry analysis and extended identity testing. *Electrophoresis* 34:1,151–1,162.

Phillips, C. 2015. Forensic genetic analysis of bio-geographical ancestry. *Forensic Sci. Int. Genetics* 18:49–65.

Pritchard J. K., M. Stephens, and P. J. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Quinto-Cortés, C. D., L. A. Arriola, G. García-Hughes et al. 2010. Genetic characterization of indigenous peoples from Oaxaca, Mexico, and its relation to linguistic and geographic isolation. *Hum. Biol.* 82:409–432.

Rangel-Villalobos, H., V. M. Martínez-Sevilla, J. Salazar-Flores et al. 2013. Forensic parameters for 15 STRs in eight Amerindian populations from the north and west of Mexico. *Forensic Sci. Int. Genetics* 7:e62–e65.

Rangel-Villalobos, H., C. D. Muñoz-Rivas, V. M. Martínez-Sevilla et al. 2014. Forensic evaluation of the AmpFℓSTR Identifiler kit in nine Mexican native populations from the pre-Columbian Mesoamerican region. *Int. J. Legal Med.* 128:467–468.

Rangel-Villalobos, H., V. M. Martínez-Sevilla, and G. Martínez-Cortés. 2016. Importance of the geographic barriers to promote gene drift and avoid pre-and post-Columbian gene flow in Mexican native groups: Evidence from forensic STR Loci. *Am. J. Phys. Anthropol.* 160:298–316.

Ricaut, F. X., C. Keyser-Tracqui, and E. Crubézy. 2005. STR-genotyping from human medieval tooth and bone samples. *Forensic Sci. Int.* 151:31–35.

Rosenberg, N. A. 2005. Algorithms for selecting informative marker panels for population assignment. *J. Comput. Biol.* 12:1,183–1,201.

Rosenberg, N. A., L. M. Li, R. Ward et al. 2003. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* 73:1,402–1,422.

Rosenberg, N. A. 2004. DISTRUCT: A program for the graphical display of population structure. *Mol. Ecol. Notes* 4:137–138.

Rosenberg, N. A., and P. P. Calabrese. 2004. Polyploid and multilocus extensions of the Wahlund inequality. *Theor. Popul. Biol.* 66:381–391.

Rubi-Castellanos, R., G. Martínez-Cortés, J. Francisco Muñoz-Valle et al. 2009. Pre-Hispanic Mesoamerican demography approximates the present-day ancestry of Mestizos throughout the territory of Mexico. *Am. J. Phys. Anthropol*. 139:284–294.

Rubi-Castellanos, R., M. Anaya-Palafox, E. Mena-Rojas et al. 2009. Genetic data of 15 autosomal STRs (Identifiler kit) of three Mexican Mestizo population samples from the States of Jalisco (west), Puebla (center), and Yucatan (southeast). *Forensic Sci. Int. Genetics* 3:e71–e76.

Rubicz, R., P. E. Melton, V. Spitsyn et al. 2010. Genetic structure of native circumpolar populations based on autosomal, mitochondrial, and Y chromosome DNA markers. *Am. J. Phys. Anthropol*. 143:62–74.

Sahoo, S., and V. K. Kashyap. 2005. Influence of language and ancestry on genetic structure of contiguous populations: A microsatellite based study on populations of Orissa. *BMC Genet.* 6:4.

Salazar-Flores, J., F. Zuñiga-Chiquette, R. Rubi-Castellanos et al. 2015. Admixture and genetic relationships of Mexican Mestizos regarding Latin American and Caribbean populations based on 13 CODIS-STRs. *Homo.* 66:44–59.

Sánchez, C., C. Barrot, M. Ortega et al. 2005. Genetic diversity of 15 STRs in Choles from northeast of Chiapas (Mexico). *J. Forensic Sci.* 50:221–223.

Schneider, P. M. 2009. Expansion of the European standard set of DNA database loci—the current situation. *Profiles in DNA* 12:6–7. http://www.promega.com/profiles/1201/1201_1206html.

Scliar, M. O., M. T. Vaintraub, P. M. Vaintraub et al. 2009. Brief communication: Admixture analysis with forensic microsatellites in Minas Gerais, Brazil: The ongoing evolution of the capital and of an African-derived community. *Am. J. Phys. Anthropol.* 139:591–595.

Seber, G. A. F., and A. J. Lee. 2003. *Linear Regression Analysis.* Hoboken, NJ: Wiley-Interscience.

Silva, N. M., L. Pereira, E. S. Poloni et al. 2012. Human neutral genetic variation and forensic STR data. *PLoS One* 7:e49666.

Simms, T. M., C. E. Rodriguez, R. Rodriguez et al. 2010. The genetic structure of populations from Haiti and Jamaica reflect divergent demographic histories. *Am. J. Phys. Anthropol.* 142: 49–66.

Snedecor, G., and W. Cochran. 1989. *Statistical Methods.* Ames, IA: Iowa State University Press.

Sprent, P. 1969. *Models in Regression and Related Topics.* London: Methuen.

Wang, J. 2003. Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* 164:747–765.

Wang, S., C. M. Lewis, M. Jakobsson et al. 2007. Genetic variation and population structure in native Americans. *PLoS Genet.* 3:e185.

Wang S., N. Ray, W. Rojas et al. 2008. Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genet.* 4:e1000037.

Zar, J. H. 1984. *Biostatistical Analysis.* Englewood Cliff, NJ: Prentice-Hall.

Table 1. Basic Statistics for Estimated Proportions of European Ancestry for the Native American, European, and Latin American Sample

Recall that it is necessary to only report statistics for only one of the two ancestry proportions calculated because of the perfect inverse relationship between the cluster proportions in the K = 2 model.

	Panel Analyzed										
Panel Statistics per Sample	1	2	3	4	5	6	7	8	9	10	Full
N											
Native American	398	387	402	393	403	382	385	376	382	397	420
European	140	143	147	150	145	154	150	132	144	139	158
Latin American	238	232	238	239	238	235	236	233	239	241	240
Median											
Native American	0.13	0.07	0.07	0.12	0.11	0.17	0.07	0.11	0.15	0.11	0.07
European	0.93	0.96	0.96	0.93	0.94	0.92	0.95	0.94	0.94	0.93	0.97
Latin American	0.83	0.83	0.82	0.78	0.77	0.80	0.61	0.81	0.83	0.86	0.56
Median Absolute Deviation											
Native American	0.09	0.05	0.05	0.09	0.08	0.12	0.04	0.08	0.11	0.08	0.06
European	0.03	0.02	0.02	0.03	0.03	0.04	0.03	0.03	0.03	0.04	0.02
Latin American	0.10	0.10	0.12	0.12	0.12	0.10	0.16	0.12	0.09	0.09	0.11
	Panel Analyzed										
Panel Statistics per Sample	1	2	3	4	5	6	7	8	9	10	Full
N											
Native American	398	387	402	393	403	382	385	376	382	397	420
European	140	143	147	150	145	154	150	132	144	139	158
Latin American	238	232	238	239	238	235	236	233	239	241	240

Median											
Native American	0.13	0.07	0.07	0.12	0.11	0.17	0.07	0.11	0.15	0.11	0.07
European	0.93	0.96	0.96	0.93	0.94	0.92	0.95	0.94	0.94	0.93	0.97
Latin American	0.83	0.83	0.82	0.78	0.77	0.80	0.61	0.81	0.83	0.86	0.56
Median Absolute Deviation											
Native American	0.09	0.05	0.05	0.09	0.08	0.12	0.04	0.08	0.11	0.08	0.06
European	0.03	0.02	0.02	0.03	0.03	0.04	0.03	0.03	0.03	0.04	0.02
Latin American	0.10	0.10	0.12	0.12	0.12	0.10	0.16	0.12	0.09	0.09	0.11

Table 2. STR Loci Included in Each of the 10 Panels

Loci 11-15 are the traditional markers included all 10 panels, while loci 1-10 were randomly selected from the pool of qualified loci from Wang et al. (2008).

CODIS Proxy Panel	1	2	3	4	5	6	7	8	9	10
Locus 1	ATCT053P_3	AGAT128_3	AAT245_17	AGAT055Z_22	AGAT113Z_13	D17S974	D2S1363	D11S1304	D13S1493	AGAT132_17
Locus 2	D10S2470	D1S1596	D19S591	D12S395	AGAT131_14	D19S254	ATAG053P_10	D15S643	D14S588	D11S1981
Locus 3	D18S535	D2S2944	D21S2052	D1S549	D15S816	D8S373	D15S659	D17S1290	GATA35_11	D13S800
Locus 4	D22S686	GGAA30H04_14	D7S1808	D2S434	D3S1768	GATA12G02_5	D1S1660	D18S542	D21S1437	GATA169F02_17
Locus 5	D7S1804	GTAT005Z_22	D9S2169	D8S2324	D4S3248	GATA135C03M_4	D1S1677	D20S481	D5S1470	D5S2505
Locus 6	GATA142C02M_20	TAGA010_5	GATA72A06_3	GATA61F04_9	D7S3061	GATA22F01_15	D20S451	D3S2387	D7S2204	D7S821
Locus 7	GATA81F06_10	TATC010P_7	TTTA040_3	GATA63B12P_15	GATA173A03_18	GATA81E09_20	D12S1064	GATA6B07_13	GATA036_18	D10S1425
Locus 8	D20S480	D10S677	AGAT115_8	D18S877	D3S1763	D14S617	D15S642	GGAA21A04_19	GGAA19G04_17	GATA165A11M_9

Locus 9	AAAT126_9	D5S1456	D22S689	GATA6F05P_22	D4S2632	D4S2368	TCTA021ZM_10	TATC046_3	GGAT2G03_3	GGAA22B10P_3
Locus 10	D13S796	GATA29_4	GATA129D03M_4	GATA23A02_2	D1S1679	D11S4463	D11S2002	D3S1744	GATA68D03_3	D10S1432
Locus 11	D13S317	D13S317	D13S317	D13S317	D13S317	D13S317	D13S317	D13S317	D13S317	D13S317
Locus 12	D16S539	D16S539	D16S539	D16S539	D16S539	D16S539	D16S539	D16S539	D16S539	D16S539
Locus 13	D19S433	D19S433	D19S433	D19S433	D19S433	D19S433	D19S433	D19S433	D19S433	D19S433
Locus 14	D7S820	D7S820	D7S820	D7S820	D7S820	D7S820	D7S820	D7S820	D7S820	D7S820
Locus 15	GATA30A08M_6	GATA30A08M_6	GATA30A08M_6	GATA30A08M_6	GATA30A08M_6	GATA30A08M_6	GATA30A08M_6	GATA30A08M_6	GATA30A08M_6	GATA30A08M_6

Table 3. *CODIS-Proxy Panels* 1-10 Linear Relationship Statistics

The number of cases for each panel, the correlation between Wang et al.'s estimates and the panel estimates on the probit scale, and the intercept, slope, and posterior standard deviation all on the probit scale.

CODIS-Proxy Panel	*N*	*r*	α	β	post. sd
1	762	0.8334	0.4383	1.4043	0.9811
2	752	0.8593	0.4002	1.2550	0.8885
3	769	0.8664	0.3518	1.2596	0.8670
4	767	0.8547	0.3773	1.4252	0.9124
5	772	0.8693	0.3463	1.3854	0.8453
6	758	0.8490	0.5185	1.5718	0.9428
7	763	0.8786	0.0835	1.3440	0.8214
8	731	0.8376	0.3815	1.3770	0.9644
9	756	0.8073	0.4965	1.5149	1.0921
10	762	0.8094	0.4746	1.4058	1.0827

Table 4. Results of Cross-Classification Analyses Performed for the European and Native American Parental Samples, and the Pooled Latino Sample

Hard-cluster assignment, or classifications, were produced by the ancestry estimates inferred from the CODIS-Proxy and Full Panels. Results were assessed by calculating the % rate of match error, letting the hard-clustering obtained from the Full Panel serve as the true or correct classification. The European-labeled column gives as % the quantity of individuals classified by the original dataset as European but who were incorrectly assigned membership in the Indigenous cluster when using the hard-cluster solutions produced from the CODIS-Proxy consensus dataset. The Indigenous-labeled column gives these same error estimates but under the opposite condition. Shaded cells correspond to instances when no classifications were made. Shaded empty cells indicate that the Full Panel did not classify any of that sample as that label, thus there was no match error to report.

Samples	Match Error (%) for CODIS Proxy Panel 7		Match Error (%) for CODIS Proxy Panel 8	
	European	Indigenous	European	Indigenous
Native American		4.21		8.15
European	1.34		2.27	
Latino	16.03	42.31	1.29	70.67

Figure 1.

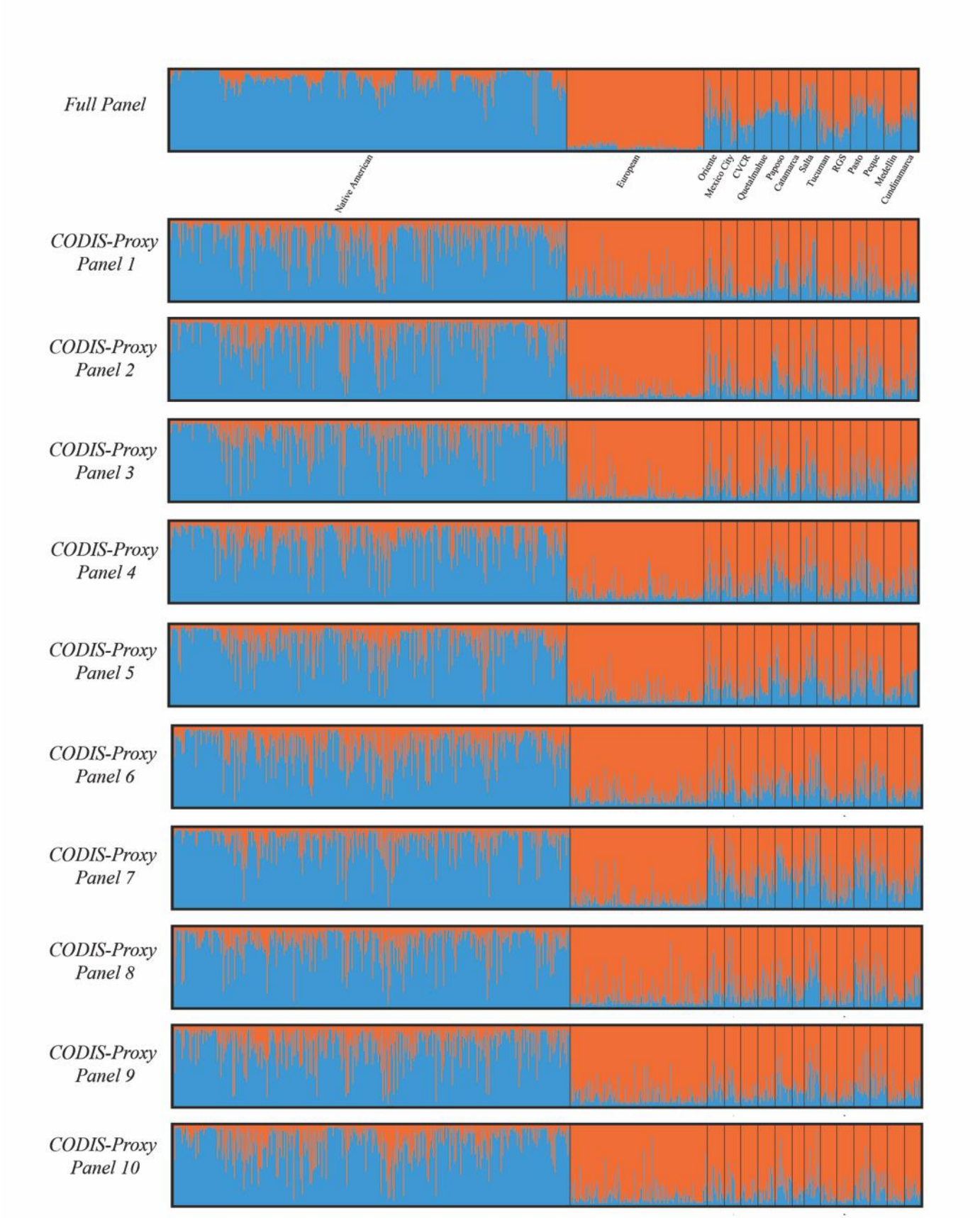

Figure 2.

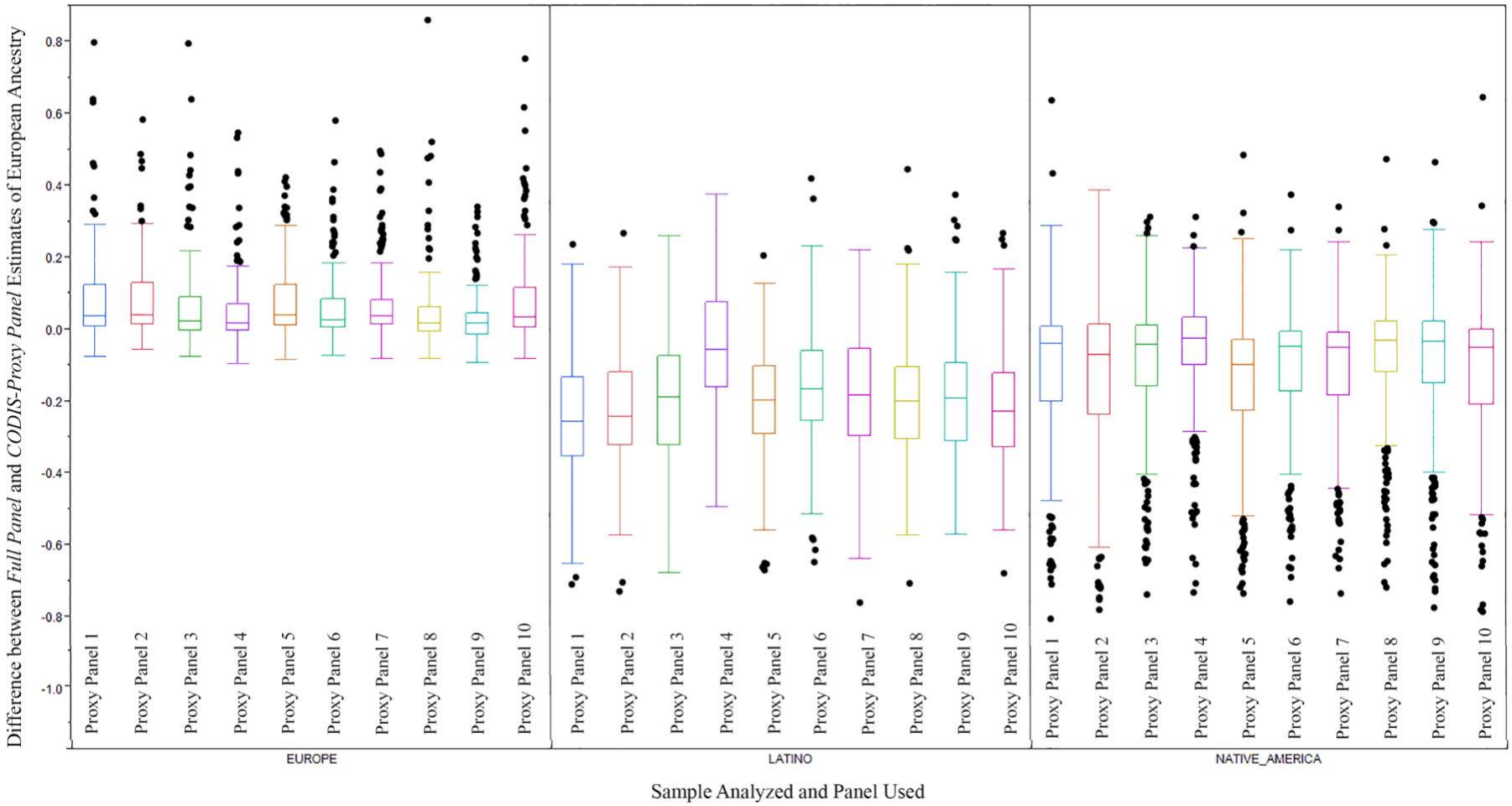
Difference between *Full Panel* and *CODIS-Proxy Panel* Estimates of European Ancestry
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
Proxy Panel 1
Proxy Panel 2
Proxy Panel 3
Proxy Panel 4
Proxy Panel 5
Proxy Panel 6
Proxy Panel 7
Proxy Panel 8
Proxy Panel 9
Proxy Panel 10
EUROPE
LATINO
NATIVE_AMERICA
Sample Analyzed and Panel Used

Figure 3.

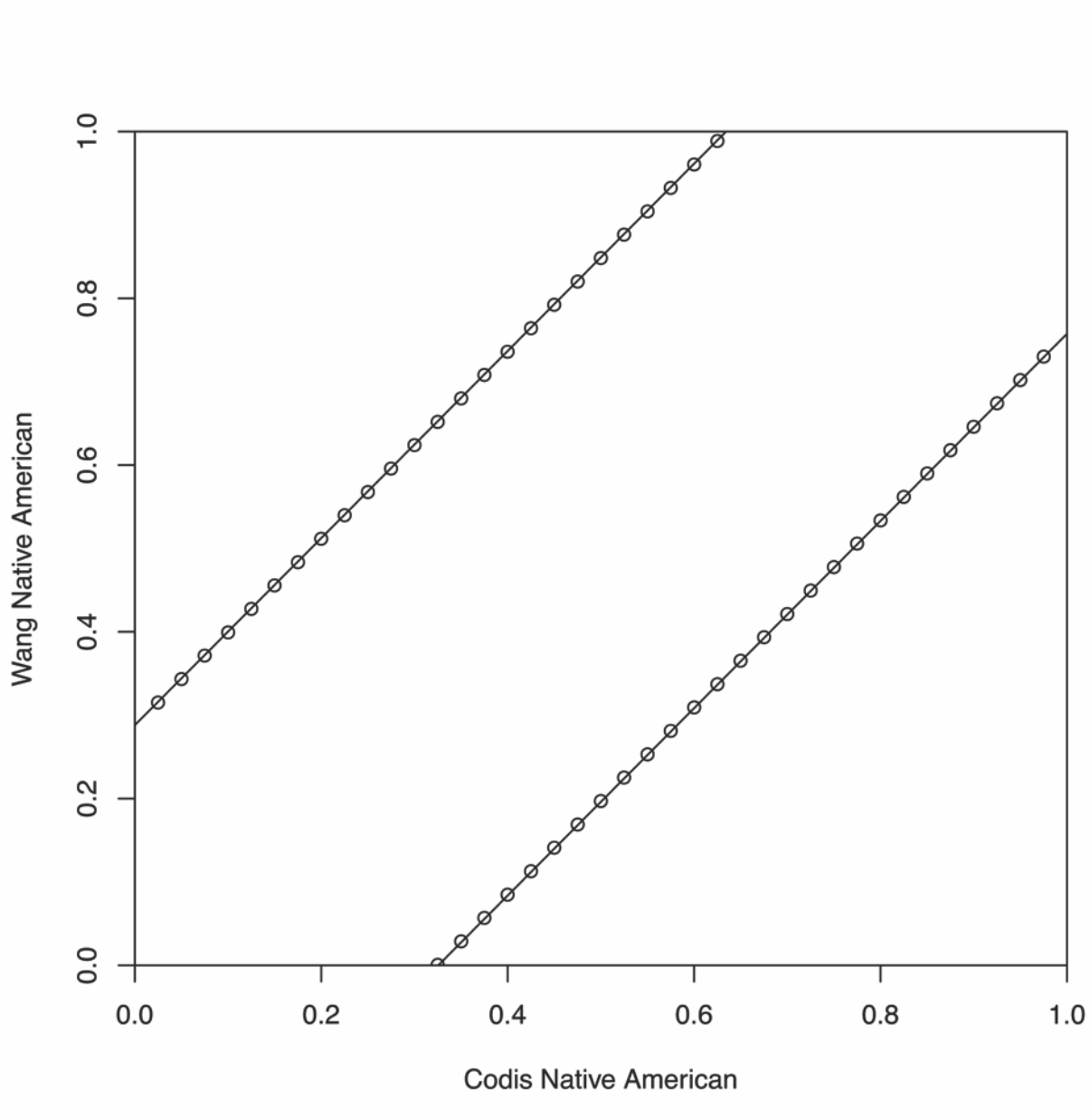

Figure 4A.

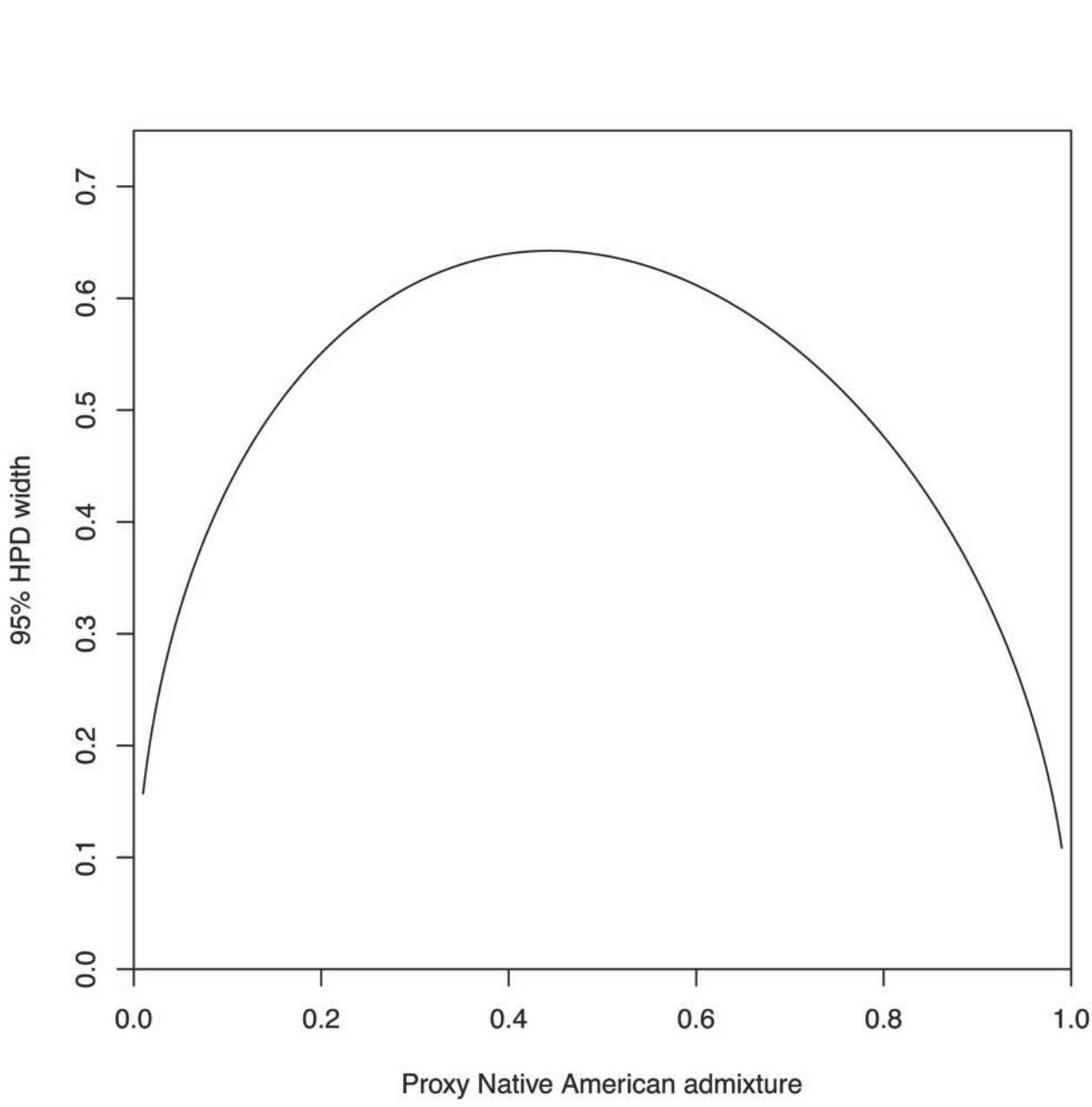

Figure 4B.

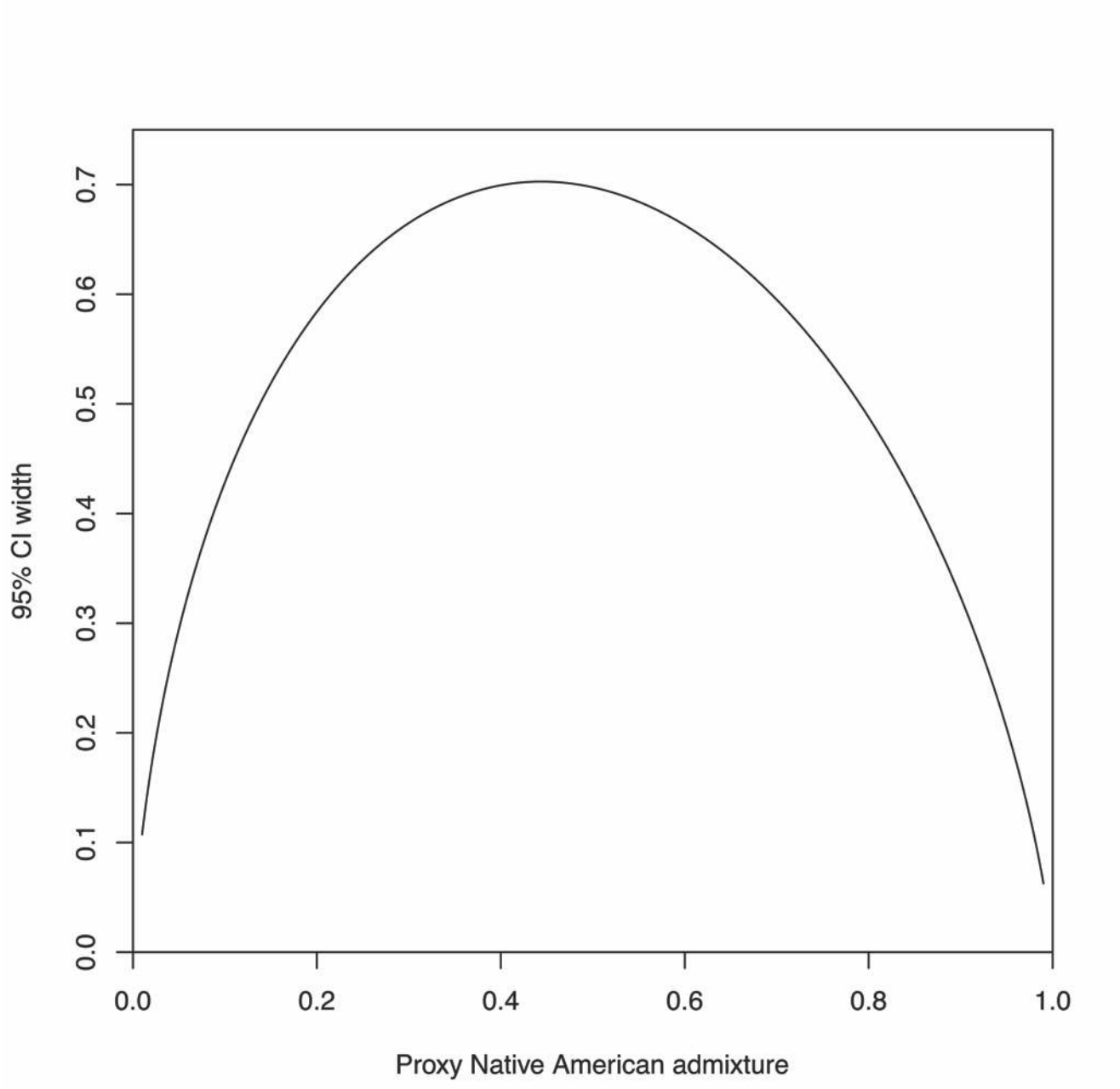

Figure 5.

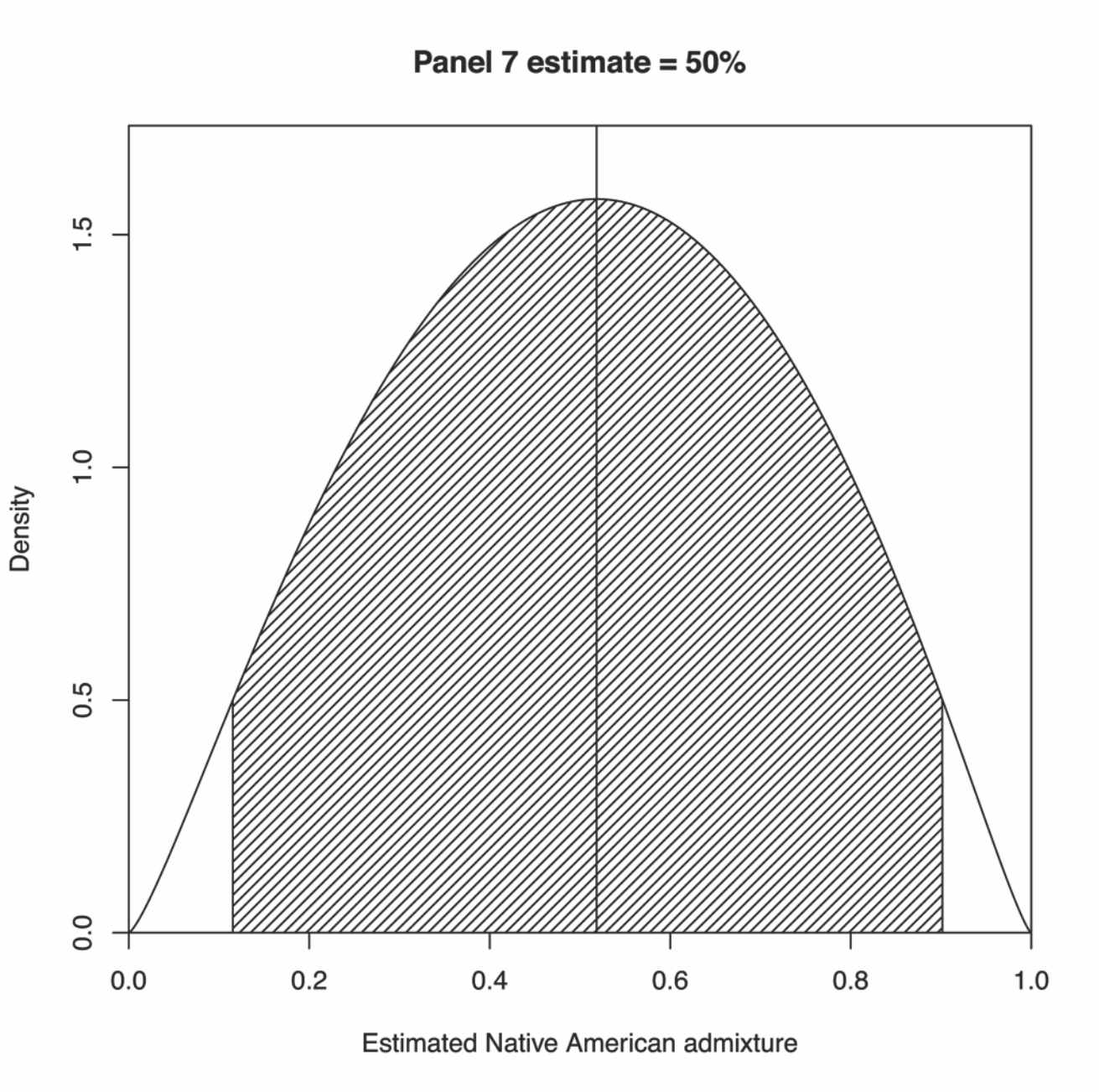

Figure Captions

Figure 1. Individual ancestry proportions for the $K = 2$ model for the European and Native American reference samples, as well as the 13 Latino samples. The *Full Panel* is provided first, followed by the *CODIS-Proxy Panels* 1-10.

Figure 2. Box plots of the individual differences in percent European ancestry estimates for the *Full Panel* with each of the 10 *CODIS-Proxy Panels*, with positive values indicating the *CODIS-Proxy Panel* underestimates European ancestry for a given individual as compared with the Full Panel, while negative values indicate the *CODIS-Proxy Panel* over-estimates European ancestry.

Figure 3. The lines are the 95% prediction intervals found by inverting the prediction intervals of y given x. The open points give the 95% prediction intervals from equation (4)

Figure 4. a) The 95% confidence interval widths for predicting the "gold standard" Native American ancestry response given the estimate of Native American ancestry from *CODIS-Proxy Panel* 7, b) the 95% highest posterior density (HPD) widths for an unknown individual's "gold standard" Native American ancestry, given the estimate of Native American ancestry from *CODIS-Proxy Panel* 7.

Figure 5. The 95% HPD for a *CODIS-Proxy Panel* estimate of 50% Native American admixture, using *CODIS-Proxy Panel* 7.